AN IMPROVED DIRECTED RANDOM WALK FRAMEWORK FOR CANCER CLASSIFICATION USING GENE EXPRESSION DATA

SEAH CHOON SEN

A thesis submitted in
fulfilment of the requirement for the award of the
Doctor of Philosophy in Information Technology

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

AUGUST 2020

For my beloved parent, family and friends who have supported me through my university years.

ACKNOWLEDGEMENT

First of all, thanks for all the accompany and sharing along the journey of doctoral, so that I am able to complete this research within the time given. Many parties have been involved throughout the completion of this research. I would like to express my sincere gratitude and respect toward my supervisor, PM. Dr. Shahreen Kasim for her continuous encouragement and guidance. Not to forget, my fiancé, Loh Yin Xia for mental support. Next, I would like to thank my beloved parent and family who has always be at my side for whatever conditions. Lastly, a very big thank you must also go to those who have always surround me with positive vibes, lend me their precious time and indirectly involved in this research.

ABSTRACT

Early diagnosis methods in cancer diagnosis studies are making great challenge as they require the involvement of different fields. Deoxyribonucleic acid (DNA) microarray analysis is one of the modern cancer diagnosis techniques used by scientists to measure the gene expression level changes in gene expression data. From the perspective of computing, an algorithm is developed to ease the diagnosis process, but the feasibility is not reliable. Numerous cancer studies have combined different machine learning techniques for the cancer diagnosis to improve the accuracy of cancer classification. This study is conducted to improve the accuracy of cancer classification by introducing an improved directed random walk (DRW) framework. This improved DRW framework is proposed to identify risk pathway while correctly predict the significant genes. It is named as significant directed walk (SDW) because of its ability to identify significant genes for cancer. In this study, six gene expression datasets are applied to study the effectiveness of the sub-algorithm, directed graph and classifier in SDW in terms of cancer prediction and cancer classification. Sub-algorithms of SDW can be further divided into data pre-processing phase, specific tuning parameter selection, weight as additional variable, and exclusion of unwanted adjacency matrix. Besides that, SDW also incorporated four directed graphs to study the usability of the directed graph. The best directed graph among the four is chosen to be part of the structure in SDW. This directed graph is the combination between KEGG pathway and PPI network and named as walker network. The experimental results showed that the combination of SDW with walker network and linear regression is the best among all. SDW achieves an accuracy of 95.03% in average which is higher by 8.97% compare to conventional DRW for all cancer datasets. This study provides a foundation for further studies and research on early diagnosis of cancer with machine learning technique. It is found that these findings would improve the early diagnosis methods of cancer classification.

ABSTRAK

Terdapat pelbagai cabaran apabila kaedah diagnosis awal dalam kajian diagnosis kanser dilaksanakan kerana ia melibatkan pelbagai bidang. Analisis microarray Asid deoksiribonukleik (DNA) merupakan salah satu teknik diagnosis kanser moden yang digunakan oleh saintis untuk mengukur tahap perubahan ekspresi gen dalam data ekspresi gen. Dari sudut perspektif pengkomputeran, algoritma dibangunkan bagi memudahkan proses diagnosis, tetapi tahap kebolehlaksanaannya diragui. Terdapat banyak kajian kanser yang telah menggabungkan teknik pembelajaran mesin yang berbeza untuk diagnosis kanser bagi meningkatkan tahap ketepatan di dalam klasifikasi kanser. Kajian ini dilakukan bagi menambah baik tahap ketepatan dalam klasifikasi kanser dengan memperkenalkan rangka kerja directed random walk (DRW) yang ditambah baik. Rangka kerja DRW yang telah ditambah baik ini dicadangkan bagi mengenal pasti laluan berisiko disamping meramalkan gen yang signifikan. Ia dinamakan sebagai significant directed walk (SDW) kerana kemampuannya dalam mengenal pasti gen yang signifikan untuk kanser. Dalam kajian ini, enam set data ekspersi gen diterapkan bagi mengkaji tahap keberkesanan sub-algoritma, graf terarah dan pengkelasan dalam SDW, dari segi ramalan kanser dan klasifikasi kanser. Sub-algoritma SDW boleh dibahagikan ke dalam fasa data pra-pemprosesan, pemilihan parameter penalaan khusus, berat sebagai pembolehubah tambahan, dan pengecualian maktrik adjacency yang tidak dikehendaki. Disamping itu, SDW turut menggabungkan empat graf terarah untuk mengkaji tahap kebolehgunaan graf terarah. Graf terarah yang terbaik di antara empat graf ini telah dipilih sebagai sebahagian daripada struktur di dalam SDW. Graf terarah ini merupakan gabungan antara laluan KEGG dan rangkaian PPI dinamakan sebagai rangkaian walker. Hasil eksperimen menunjukkan bahawan kombinasi antara SDW dengan rangkaian walker dan regresi linear adalah yang terbaik di antara semua. SDW mencapai tahap ketepatan sebanyak 95.03% secara purata, dimana ia lebih tinggi sebanyak 8.97% berbanding DRW konvensional bagi semua set data kanser. Kajian ini memberikan asas kepada kajian

lanjutan serta penyelidikan mengenai diagnosis awal kanser dengan teknik pembelajaran mesin. Hasil dapatan kajian ini akan menambah baik kaedah diagnosisawal klasifikasi kanser.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AUC	-	Area under receiver operating characteristics curve
BGLR	-	Bayesian Generalized Linear Regression
C-index	-	Concordance index
CNN	-	Convolutional Neural Network
CRW	-	Correlated random walks
DEGS	-	Differentially expressed genes
DNA	-	Deoxyribonucleic acid
DRW	-	Directed Random Walk
ECM	-	extracellular matrix
eQTL	-	Expression quantitative trait loci
ESCC	-	Esophageal squamous cell carcinoma
FN	-	False negative
FP	-	False positive
GBM	-	Glioblastoma multiforme
GCRMA	-	Gene Chip Robust Multiarray Averaging
GEO	-	Gene Expression Omnibus
GSEA	-	Gene set enrichment analysis
HPRD	-	Human protein reference database
ID	-	Genes identity
iDRW	-	Integrative directed random walk

IPA	-	Ingenuity Pathway Analysis
KEGG	-	Kyoto Encyclopedia of Genes and Genomes
lncRNA	-	Long non-coding Ribonucleic acid
MCMC	-	Markov chain Monte Carlo
MERW	-	Maximal entropy random walk
mRNA	-	Messenger Ribonucleic acid
NCBI	-	National Centre for Biotechnology Information
NGS	-	Next generation sequencing
NPR	-	Nonparametric Pathway based Regression
PAGE	-	Parametric Analysis of Gene Set Enrichment
PARADIGM	-	Pathway Recognition Algorithm using Data Integration on Genomic Models
PPI	-	Protein-protein interaction
REC	-	Recall
RF	-	Random forests
RNA-SEQ	-	Ribonucleic acid sequencing
RPCLR	-	Random-Penalized Conditional Logistic Regression
RRSF	-	Reweight random survival forest
sDRW	-	Significant directed random walk
SDW	-	Significant directed walk
SIRW	-	Self-interacting random walk
TN	-	True negative
TP	-	True positive
TPR	-	True positive rate

LIST OF APPENDICES

CHAPTER 1

OVERVIEW

1.1 Background

Deoxyribonucleic acid (DNA) microarray analysis is one of the modern technologies used by scientists to measure the gene expression level changes in gene expression data. Microarray analysis involves breaking a cell, isolating its genetic contents, identifying all the genes that are turned on in that particular cell and generating a list of those genes data (DNA Microarray) (Subat *et al.,* 2018). Even though gene expression data is full of genetic information, it is not enough to identify genetic diseases such as cancer, Hemochromatosis, Huntington's disease, Turner syndrome, and Alzheimer's disease (Velsher, 2003). A single genetic dataset is found to be insufficient to predict, as well as classify, the cancer in the human body perfectly (Zhang *et al.*, 2019). Hence, the integration between pathway data and gene expression data in the search process can provide a better understanding in biological processes for disease detection as well as cancer classification (Ong *et al.,* 2011). Biologists and researchers need an accurate classification tool and lists of potential genetic disorder's gene during the disease diagnosis process (Tekade & Rajeswari, 2018).

The result of microarray analysis, gene expression data are characterized by a large number of genes (variables) from a small number of patients infected by the same disease. This leads to a high degree of multi-collinearity (Bhattacharjee & Vishwakarma, 2019). This problem also named as high dimension small sample size (Liu *et al.,* 2015). Similar issue happens in pathway data. Normally, pathway data usually gathered from the biological literature and defined from the biological context free. The genes in a pathway data are responsible to a specific cellular process (Misman *et al.,* 2011). Some irrelevant genes present in the gene expression data degrade the

data quality (Ibrahim *et al.*, 2011). The mutual variable between these two datasets can help in merging the data and solve the high dimension small sample size problem (Paszkiewicz & Studholme, 2011).

From the biological literature (Wang *et al*., 2008), cancer is one of the multifactorial genetic disorder diseases which is caused by a combination of environmental factors and mutations in multiple genes where the gene is tied to the biological pathway to generate the needed protein in human body. Gene expression data have all genes from selected sample cell (set), while pathway data only provide certain number of genes (subset). Some uninformative genes may include within the gene expression data due to the high degree of gene variables. Hence, only small number of genes in the gene expression data are responsible to a specific cellular process. For example, different genes that influence breast cancer susceptibility have been found on chromosomes 6, 11, 13, 14, 15, 17, and 22 (Stöppler, 2019).

Many researches related to search algorithm in cancer classification had been studied to further enhance the ability of the algorithms as well as improve the performance (Polat & Güneş, 2009). Examples of these algorithms are feature selection (Kang *et al.*, 2019), convolutional neural network (CNN) (Fan *et al.*, 2019), and naive bayes (Wood *et al.*, 2019) and random walk (Buraczewski & Dyszewski, 2018), of which, are commonly use in data classification. However, the greatest weakness among the search algorithm is the blindly direct and search among all data (Rami-Porta & Goldstraw 2010). The direction guide applied in biased random walk has great potential to further develop and overcome the weakness of the cancer classification (Codling *et al.*, 2008).

Directed random walk (DRW), as one of the famous approaches specialized for cancer classification, was developed from biased random walk. It was developed to infer reproducible pathway activities and robust disease classification (Liu *et al*., 2013). Not only it is computational complex, but there is some limitation in the algorithm. For example, overfitting in DRW occurs due to the restart probability. When the restart probability is high, overfitting will be occurred, and it will lead to poor discriminating power. Discriminating power of the algorithm shows the reproducible power and the robustness of the directed pathways whether the pathways are significantly differentially expressed or not (Liu *et al.*, 2013). Besides, the ability of DRW is also limited by the poor reference data pool, which only consisted of one type reference data.

In a nutshell, an improved directed random walk framework to identify more cancerous genes and higher accuracy in cancer classification of gene expression data is proposed to tackle the problems. More pathway data are used as the directed graph for biased random walk while gene expression data will go through the biased random walk process with the directed graph. Gene expression values are utilized as one of the variables as it is one of the key values from gene expression data. Risk pathway, also known as potential genetic disorder pathway, are selected, and further extracted in order to identify the risk genes. We expect that the performance of the improved algorithm will be superior as compared to other previous related algorithms, including the conventional version and other improved version of directed random walk in terms of classification accuracy and the number of cancerous genes being identified.

1.2 Problem Statements

Random Walk has gained the attention of researchers due to its flexibility in multiple fields such as bio-technology (Codling *et al.*, 2008), finance (Misra & Chaurasia, 2018), as well as data analytic (Dudziński *et al.*, 2019). In 2013, directed random walk (DRW) was presented by Liu as a specialized algorithm for cancer classification (Liu *et al.*, 2013). This algorithm had proved the accuracy of cancer classification by iterating the random walk process with the directed graph that was built by Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway data (Liu *et al.*, 2013).

There are some limitations such as unused important attributes on the conventional directed random walk approach which lead to low sensitivity and limited accuracy of cancer classification (Wang & Liu, 2018; Kim *et al.*, 2018; Ning *et al.*, 2019; Kim *et al.*, 2019). In order to predict the cancerous gene correctly and classify the gene accurately, three main related factors are identified in this work as following:

- Feasibility of DRW algorithm to control search direction towards cancerous gene during the diagnosis process.
- The usage of limited parameters in DRW algorithm to analyse cancerous gene during the diagnosis process.
- Limited data source as reference in directed graph.

The first factor related to the feasibility of search direction in directed random walk approach (Kim *et al.*, 2018). Feasibility of search direction can be described as

computational flow and arrangement of order in data pre-processing. The DRW approach did not perfectly match with multiple types of cancer datasets (Kim *et al.*, 2019). The randomness of DRW is uncontrollable and did not fulfil the biological rules of protein formation (Huang, 2009). This research opened up ideas to apply suitable flows in DRW for a better search direction.

The second factor related to the usage of parameters and data analysis in directed random walk approach (Wang & Liu, 2018). The usability of parameters and data analysis are about the usage of variables in enhancing the ability of data analysis in DRW algorithm for cancer detection purposes (Ning *et al.*, 2019). Variables such as restart probability, weight, and adjacency matrix are useful in assisting the algorithm to identify the cancerous genes. This research opened up ideas to apply flexible, and meaningful variables in DRW for a better cancerous gene prediction.

The third factor, inspired by the limitation of directed graph where only 1 set of KEGG data group is applied, named as global pathway network (Wang & Liu, 2018). The limitation of this class of global pathway network is their ignorance of interactions between genes and proteins because neither network topology nor dynamics is considered (Zhang *et al.*, 2019). This research came up with the idea to enrich the biological databases with more sample genes and protein information on the directed graph.

Hence, an idea came up to improve the DRW framework as well as algorithm and directed graph which mainly focus on enhancement of the cancer classification accuracy and sensitivity of cancer gene prediction.

1.3 Research Questions

There are some questions that had been asked to clearly identify the research gap. These questions are identified in this work as following:

- What are the main issues of DRW algorithm that affect its ability to accurately control the search direction towards cancerous genes?
- What are the variables that can be considered in improving the data analysis in DRW algorithm?
- What kind of data sources that can be used to enrich the data pool with more pathway activities?

1.4 Research Objectives

This research aimed to enhance and improve the directed random walk framework to increase the accuracy of cancer classification. The objectives of this research are:

- To propose significant directed walk (SDW) as an improved directed random walk algorithm that enhances the search direction and data analysis.
- To propose walker network as a larger directed graph in order to enrich the data pool with more pathway activities.
- To evaluate the proposed significant directed walk with existing classifier in term of accuracy.

1.5 Significant of the Study

The main contributions of this study are in the field of machine learning where the details of these contributions are identified in this work.

An improved directed random walk algorithm in field of cancer classification is able to perform higher accuracy and better sensitivity in cancer prediction. Data pre-processing phase is introduced as the first step for data cleaning purposes. Sub-algorithms, such as specific tuning parameter selection, weight as additional variable, excluded unwanted adjacency matrix, were introduced to enhance the search direction and data analysis. This algorithm achieved higher accuracy as compared to conventional DRW for all cancer datasets. This is rational because SDW is more effective, and feasible for cancer classification as well as cancer prediction compare to the other methods.

A larger directed graph with the combination of two different type of genes' co-relative pathway dataset to enrich the data pool. Combination between KEGG pathway network and PPI network are named as walker network. The walker network is able to enhance the prediction performance with more risk pathways and significant genes compare to other directed graphs. This is rational because the benefit of the biological pathway in KEGG are combined with the strength of protein generation sequences in PPI network.

1.6 Research Scope

This research only focuses on improving the directed random walk (Liu *et al.,* 2013) in term of cancer classification through R programming. R studio is used as development platform for the proposed approach. Gene expression data is used as input data while additional pathway data is used as reference data. Six cancerous datasets, lung, liver, thyroid, stomach, kidney, and breast are chosen as the only input dataset. The pathway data that used to form directed graph are KEGG pathway and protein-protein interaction (PPI) network. The performance measurement is measured by the predicted number of cancerous gene and the accuracy of cancer classification via area under receiver operating characteristics curve (AUC).

1.7 Thesis Organization

This chapter gave the overview with motivation for improvement of directed random walk in cancer classification, problem statements and research objectives with the scope of this research.

Chapter 2 explained the fundamental of basic theory of cancer classification with microarray data analysis and a survey of literature. This chapter discussed the algorithm for cancer classification used in microarray data analysis. The review of comparative published literature and evaluation criteria is also discussed in this chapter.

Chapter 3 described the research methodology. In this chapter, an improved algorithm is proposed which shows how the research will be conducted. The improvement comprised of three phases. First phase presented the significant directed walk (SDW) framework as well as the sub-algorithm with mathematical formula while second phase proposed the new larger directed graph for improved directed random walk. Lastly, the third phase improved the usage of data by introducing a match's classifier. Moreover, this chapter also explained the experimental setup in R programming from the perspective of implementation tools.

Chapter 4 analysed the conventional directed random walk and proposed the SDW framework as well as introduced a new data pre-processing method to provide a clean dataset. The improved sub-algorithm in SDW is introduced to improve the approach from performance perspective such as accuracy of cancer classification and

REFERENCES

Aggarwal, D., and Gupta, D., (2019). Review of Decision Tree Based Classification Algorithms in Medical Data. *International Journal of Computer Sciences and Engineering*, 7(5), 2347-2693

Ahmed, R., Baali, I., Erten, C., Hoxha, E., & Kazan, H. (2019). MEXCOwalk: Mutual Exclusion and Coverage Based Random Walk to Identify Cancer Modules. Bioinformatics,

Alamgir, M., & Luxburg, U. V. (2010). Multi-agent Random Walks for Local Clustering on Graphs. *2010 IEEE International Conference on Data Mining*.

Aldous and Fill. (2014), Reversible Markov Chains and Random Walks on Graphs. Retrieved February 15, 2017, from http://www.stat.berkeley.edu/~aldous/RWG/book.html

Aldous, D. and Shepp L. (1987). The Least Variable Phase Type Distribution Is Erlang. Communications in Statistics. *Stochastic Models, 3(3)*, 467–473.,

Al-Rajab, M., & Lu, J. (2014). Algorithms Implemented for Cancer Gene Searching and Classifications. *The 10th International Symposium on Bioinformatics Research and Applications (ISBRA2014)*

Andrea Schmidt (2017). Random Walk. Retrieved February 19, 2017, from http://www.mit.edu/~kardar/teaching/projects/chemotaxis(AndreaSchmidt)/random.htm

Attanayake A, Jayasundara D, Peiris T. (2016). An Application Of 5-Fold Cross Validation on A Binary Logistic Regression Model. *Advances and Applications in Statistics. 49(6)*, 443-451.

Balasubramanian R. (2012). Package 'RPCLR' [Internet]. Cran.r-project.org. Retrieved September 2018, from: https://cran.r-project.org/web/packages/RPCLR/RPCLR.pdf

Banupriya, C. S. (2016). A Survey on Destination Prediction Using Trajectory Data Mining Technique. *International Journal of Engineering And Computer Science*.

Bates, D., Maechler, M., & Davis, T. A. (2019). Sparse and Dense Matrix Classes and Methods [R package Matrix version 1.2-18]. Retrieved from https://cran.rproject.org/web/packages /Matrix/index.html

Bender, M. A., & Ron, D. (2002). Testing properties of directed graphs: acyclicity and connectivity. *Random Structures and Algorithms, 20(2),* 184-205.

Benhamou, S. (2014). Of scales and stationarity in animal movements. *Ecology Letters, 17,* 261–72.

Bhattacharjee, A., & Vishwakarma, G. K. (2019). Time-course data prediction for repeatedly measured gene expression. *International Journal of Biomathematics, 12(04),* 1950033.

Brandt, P. A. Van Den, (2000). Pooled Analysis of Prospective Cohort Studies on Height, Weight, and Breast Cancer Risk. *American Journal of Epidemiology, 152(6),* 514–527.

Buraczewski, D., & Dyszewski, P. (2018). Precise large deviations for random walk in random environment. *Electronic Journal of Probability, 23(0).*

Cai H, Ruan P, Ng M, Akutsu T. (2014). Feature weight estimation for gene selection: a local hyperlinear learning approach. *BMC Bioinformatics.* 15:70.

Campos G, Pataki A, Pérez P. (2013) The BGLR (Bayesian Generalized Linear Regression) R-Package [Internet]. Bglr.r-forge.r-project.org. [cited 4 September 2018]. Available from: http://bglr.r-forge.r-project.org/BGLR-tutorial.pdf

Chen, A., & Renshaw, E. (1994). The general correlated random walk. *Journal of Applied Probability, 31(4),* 869-884.

Choudum, S. (1986). A simple proof of the Erdos-Gallai theorem on graph sequences. *Bulletin of the Australian Mathematical Society, 33(01),* 67.

Chua, M. E., Tanseco, P. P., Mendoza, J. S., Castillo, J. C., Morales, M. L., & Luna, S. L. (2015). Configuration and validation of a novel prostate disease nomogram predicting prostate biopsy outcome: A prospective study correlating clinical indicators among Filipino adult males with elevated PSA level. *Asian Journal of Urology, 2(2),* 114–122.

Cocktell, S. (2011). Gene Set Expression Analysis. Retrieved May 17, 2018, from http://bioinformatics.knowledgeblog.org/2011/06/20/gene-set-enrichment-analysis/

Codling, E. A., Plank, M. J., & Benhamou, S. (2008). Random walk models in biology. *Journal of The Royal Society Interface, 5(25),* 813-834.

Coudray, N., Moreira, A. L., Sakellaropoulos, T., Fenyö, D., Razavian, N., & Tsirigos, A. (2017). Classification and Mutation Prediction from Non-Small Cell Lung Cancer Histopathology Images using Deep Learning. *Nature Medicine, 24 (10),* 1559 -1569.

Creighton, C. J., & Rae, J. M. (2006). When will tumor gene expression profiling be incorporated into clinical breast cancer decision making? *Breast Cancer Research, 8(4).*

Cristianini, N. & Shawe-Taylor, J. (1999). An introduction to support vector machines. Cambridge, *MA: Cambridge University Press*

Cupertino, T. H., Guimarães Carneiro, M., Zheng, Q., Zhang, J., & Zhao, L. (2018). A scheme for high level data classification using random walk and network measures. *Expert Systems with Applications, 92,* 289-303.

D'Errico, M., Rinaldis, E. D., Blasi, M. F., Viti, V., Falchetti, M., Calcagnile, A., Sera, F., Saieva, C., Ottini, L., Palli, D., Palombo, F., Giuliani, A., & Dogliotti, E. (2009). Genome-wide expression profile of sporadic gastric cancers with microsatellite instability. *European Journal of Cancer, 45(3),* 461–469.

Dai, J. Y., LeBlanc, M., Goodman, P. J., Lucia, M. S., Thompson, I. M., & Tangen, C. M. (2019). Case-only Methods Identified Genetic Loci Predicting a Subgroup of Men with Reduced Risk of High-grade Prostate Cancer by Finasteride. *Cancer Prevention Research, 12(2),* 113-120.

Dai, Y., Guo, L., Li, M. and Chen, Y. (2012). Microarray *Я* US: a user-friendly graphical interface to Bioconductor tools that enables accurate microarray data analysis and expedites comprehensive functional analysis of microarray results. *BMC Research Note. 5(1),* 282.

Dalgliesh, G. L., Furge, K., Greenman, C., Chen, L., Bignell, G., Butler, A., Davies, H., Edkins, S., Hardy, C., Latimer, C., Teague, J., Andrews, J., Barthorpe, S., Beare, D., Buck, G., Campbell, P. J., Forbes, S., Jia, M., Jones, D., Knott, H., Kok, C., Y., Lau, K. W., Leroy, C., Lin, M.-L., McBride, D. J., Maddison, M., Maguire, S., McLay, K., Menzies, A., Mironenko, T., Mulderrig, L., Mudie, L., O'Meara, S., Pleasance, E., Rajasingham, A., Shepherd, R., Smith, R., Stebbings, L., Stephens, P., Tang, G., Tarpey, P. S., Turrell, K., Dykema, K. J., Khoo, S. K., Petillo, D., Wondergem, B., Anema, J., Kahnoski, R. J., Teh, B. T., Stratton, M. R., & Futreal, P. A. (2010). Systematic sequencing of renal

carcinoma reveals inactivation of histone modifying genes. *Nature, 463(7279),* 360–363.

Dao, H., & Jin, H. (2015). Data Preprocessing and Classification for Taproot site data sets of Panax notoginseng. *Proceedings of the 2nd International Conference on Modelling, Identification and Control*.

Deepa S. Deukar and R.R. Deshmukh. (2016). Data Mining Classification. Imperial *Journal of Interdisciplinary Research, 2(4),* 2454-1362

DNA Microarray. (n.d.), Retrieved May 12, 2017, from http://learn.genetics.utah.edu/content/labs/microarray/

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., Georgescu, C., & Romero, R. (2007). A systems biology approach for pathway level analysis. *Genome Research, 17(10),* 1537–1545.

Dudziński, M., Furmańczyk, K., & Orłowski, A. (2019). Some Proposal of The Test For A Random Walk Detection And Its Application In The Stock Market Data Analysis. *Metody Ilościowe w Badaniach Ekonomicznych, 19(4),* 339–346.

Edelstein, H., A. (1999). Introduction to data mining and knowledge discovery (3rd ed). *Potomac*, MD: Two Crows Corp.

Edgar, R. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research, 30(1),* 207–210.

Efroni, S., Schaefer, C. F., & Buetow, K. H. (2007). Identification of Key Processes Underlying Cancer Phenotypes Using Biologic Pathway Analysis. *PLoS ONE*, 2(5).

Fan, K., Wen, S., & Deng, Z. (2019). Deep Learning for Detecting Breast Cancer Metastases on WSI. *Innovation in Medicine and Healthcare Systems, and Multimedia Smart Innovation, Systems and Technologies,* 137-145.

Fox, E. J., & Reid-Bayliss, K. S. (2014). Accuracy of Next Generation Sequencing Platforms. *Journal of Next Generation Sequencing & Applications, 01(01).*

Fruzangohar, M., Ebrahimie, E., & Adelson, D. L. (2014). Application of Global Transcriptome Data in Gene Ontology Classification and Construction of a Gene Ontology Interaction Network. *bioRxiv.*

Gao X, Chen F, Song F, Jin Z. (2009). Influence of feature weight on text categorization performance of Bayesian classifier. *Journal of Computer Applications. 28(12)*:3080-3083.

Gibbons, F., Roth, F. (2002). Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation. *Genome Research. 12(10)*: 1574-1581.

Gold, D. L., Miecznikowski, J. C., & Liu, S. (2009). Error control variability in pathway-based microarray analysis. *Bioinformatics, 25(17),* 2216–2221.

Graudenzi, A. (2017) Pathway-based classification of breast cancer subtypes. *Frontiers in Bioscience. 22(10*):1697-1712.

Guo, Z., Zhang, T., Li, X., Wang, Q., Xu, J., Yu, H., Zhu, J., Wang, H., Wang, C., Topol, E. J., Wang, Q., & Rao, S. (2005). Towards precise classification of cancers based on robust gene functional expression profiles, *BMC Bioinformatics, 6(1),* 58.

Guyon Isabelle, Weston J. & Barnhill S. (2001) Gene Selection for Cancer Classification using Support Vectore Machines. *Machine Learning, 46,* 389-422

Guzeldemir-Akcakanat, E., Sunnetci-Akkoyunlu, D., Orucguney, B., Cine, N., Kan, B., Yılmaz, E. B., Gümüşlü, E., & Savli, H. (2016). Gene-Expression Profiles in Generalized Aggressive Periodontitis: A Gene Network-Based Microarray Analysis. *Journal of Periodontology, 87(1),* 58–65.

H. H. Jeong, S. Y. Kim, K. Wee, and K. A. Sohn. (2015). Investigating the utility of clinical outcome-guided mutual information network in network-based Cox regression. *BMC systems biology, 9*(Suppl 1), S8.

Haar, L., Anding, K., Trambitckii, K., & Notni, G. (2019). Comparison between Supervised and Unsupervised Feature Selection Methods. *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*.

Han, J., Kamber, M. (2000). Data mining: Concepts and Techniques. *New York: Morgan-Kaufman*

Hijazi, H., & Chan, C. (2013). A Classification Framework Applied to Cancer Gene Expression Profiles. *Journal of Healthcare Engineering, 4(2),* 255-284.

Huang, L. (2009). An integrated method for cancer classification and rule extraction from microarray data. *Journal of Biomedical Science, 16(1),* 25.

Ibrahim, M. A, Jassim, S., Cawthorne, M. A., & Langlands, K. (2011). A Pathway-based Gene Selection for Disease Classification. *IEEE*

Jadamba E, Shin M. (2014). A novel approach to significant pathway identification using pathway interaction network from PPI data. *BioChip Journal. 8(1):*22-27.

Jason Brownlee (2016), Logistic Regression for Machine Learning. Retrieved May 19, 2018, from http://machinelearningmastery.com/logistic-regression-for-machine-learning/

Jogia, G., Tronser, T., Popova, A., & Levkin, P. (2016). Droplet Microarray Based on Superhydrophobic-Superhydrophilic Patterns for Single Cell Analysis. *Microarrays, 5(4),* 28.

Johannes, M., Frohlich, H., Sultmann, H., & Beissbarth, T. (2011). pathClass: An R-package for integration of pathway knowledge into support vector machines for biomarker discovery. *Bioinformatics, 27(10),* 1442–1443.

Jones, J. (2005). Gene Signatures of Progression and Metastasis in Renal Cell Cancer. *Clinical Cancer Research, 11(16),* 5730–5739.

Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res., 28,* 27–30.

Kang, C., Huo, Y., Xin, L., Tian, B., & Yu, B. (2019). Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. *Journal of Theoretical Biology, 463,* 77–91.

Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support Vector Machines in R. *Journal of Statistical Software, 15(9),* 1-28.

Kegg Pathway: Leukocyte transendothelial migration - Homo sapiens (human) (2017). Genome.jp. Retrieved May 28, 2018, from http://www.genome.jp/kegg-bin/show_pathway?hsa04670

Kim S. Y., Volsky D. Y., (2005). PAGE: Parametric Analysis of Gene Set Enrichment, *Columbia University Academic Commons.*

Kim, J. H., Karnovsky, A., Mahavisno, V., Weymouth, T., Pande, M., Dolinoy, D. C., Sartor, M. A. (2012). LRpath analysis reveals common pathways dysregulated via DNA methylation across cancer types. *BMC Genomics, 13(1),* 526.

Kim, S. Y., Jeong, H.-H., Kim, J., Moon, J.-H., & Sohn, K.-A. (2019). Robust pathway-based multi-omics data integration using directed random walks for survival prediction in multiple cancer studies. *Biology Direct, 14(1).*

Kim, S. Y., Kim, T. R., Jeong, H., & Sohn, K. (2018). Integrative pathway-based survival prediction utilizing the interaction between gene expression and DNA methylation in breast cancer. *BMC Medical Genomics, 11(S3).*

Kittas, A., Delobelle, A., Schmitt, S., Breuhahn, K., Guziolowski, C., & Grabe, N. (2015). Directed random walks and constraint programming reveal active pathways in hepatocyte growth factor signaling. *FEBS Journal, 283(2)*, 350–360.

Köhler, S., Bauer, S., Horn, D., & Robinson, P. N. (2008). Walking the Interactome for Prioritization of Candidate Disease Genes. *The American Journal of Human Genetics, 82(4)*, 949–958.

Komurov, K., White, M. A., & Ram, P. T. (2010). Use of Data-Biased Random Walks on Graphs for the Retrieval of Context-Specific Networks from Genomic Data. *PLoS Computational Biology, 6(8).*

Landi, M. T., Dracheva, T., Rotunno, M., Figueroa, J. D., Liu, H., Dasgupta, A., Mann, F. E., Fukuoka, J., Hames, M., Bergen, A. W., Murphy, S. E., Yang, P., Pesatori, A. C., Consonni, D., Bertazzi, P. A., Wacholder, S., Shih, J. H. Caporaso N. E., & Jen, J. (2008). Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival. *PLoS ONE, 3(2).*

Lawler, G. F., & Limic, V. (1996). Loop-erased random walk. Random Walk: A Modern Introduction, 307-325.

Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T., & Lee, D. (2008). Inferring Pathway Activity toward Precise Disease Classification. *PLoS Computational Biology, 4(11).*

Li, C., Li, X., Miao, Y., Wang, Q., Jiang, W., Xu, C., Jing Li, Han, J., Zhang, F., Gong, B., & Xu, L. (2009). SubpathwayMiner: a software package for flexible identification of pathways. *Nucleic Acids Research, 37(19).*

Li, J., Zhu, J., & Zhang, B. (2016). Discriminative Deep Random Walk for Network Classification. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).*

Li, Y., Li, G., Wang, F., Wu, X., Wu, Z., Wang, J., Zhang, C., He, J., Wang, H., & Wang, S. (2019). Integrated Analysis of LncRNA-mRNA Coexpression in the Extracellular Matrix of Developing Deciduous Teeth in Miniature Pigs. *BioMed Research International*, 1–9.

Lin, C., Lin, R., Chen, T., Zigler, C., Wei, Y., & Christiani, D. C. (2019). A global perspective on coal-fired power plants and burden of lung cancer. *Environmental Health, 18(1).*

Lin, Y., & Zhang, Z. (2014). Mean first-passage time for maximal-entropy random walks in complex networks. *Scientific Reports, 4(1).*

Liu, J., Xu, Y., Zheng, C., Kong, H., & Lai, Z. (2015). RPCA-Based Tumor Classification Using Gene Expression Data. *IEEE/ACM Transactions Computer Biological and Bioinformatics. 12(4),* 964-970.

Liu, W., Li, C., Xu, Y., Yang, H., Yao, Q., Han, J., Shang, D., Zhang, C., Su, F., Li, X., Xiao, Y., Zhang, F., Dai, M., & Li, X. (2013). Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. *Bioinformatics, 29(17),* 2169–2177.

Luecken, M. D., Page, M. J., Crosby, A. J., Mason, S., Reinert, G., & Deane, C. M. (2017). CommWalker: Correctly evaluating modules in molecular networks in light of annotation bias. *Bioinformatics, 34(6),* 994-1000.

Maeda, N. (2019). New Era of Genome-Wide Gene Expression Analysis. *Cap-Analysis Gene Expression (Cage),* 61-78.

Malachowicz, M., & Wenne, R. (2019). Microarray analysis of gene expression of Atlantic cod from different Baltic Sea regions: Adaptation to salinity. *Marine Genomics.*

Margolin, A. A., & Califano, A. (2007). Theory and Limitations of Genetic Network Inference from Microarray Data. *Annals of the New York Academy of Sciences, 1115(1),* 51-72.

Meghanathan N. (2015). Exploiting the Discriminating Power of the Eigenvector Centrality Measure to Detect Graph Isomorphism. *International Journal in Foundations of Computer Science & Technology, 5(6):*01-13.

Meštrović, T. (2019). Types of Microarray. Retrieved June 17, 2019, from https://www.news-medical.net/life-sciences/Types-of-Microarray.aspx

Michy Alice. (2015). How to perform a Logistic Regression in R. Retrieved May 25, 2017, from https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T., & Bergh, J. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences, 102(38),* 13550–13555.

Misman, M. F., Mohamad, M. S., Deris, S., Abdullah, A., & Hashim, S. Z. (2011). An improved hybrid of SVM and SCAD for pathway analysis. *Bioinformation, 7(4),* 169-175.

Misra, P., & Chaurasia, S. (2018). Financial Market Predictions Generative Vs Discriminative Methods. *International Journal of Computer Sciences and Engineering, 6(7),* 1373–1378.

Möhlendick, B., & Stoecklein, N. H. (2014). Analysis of Copy-Number Alterations in Single Cells Using Microarray-Based Comparative Genomic Hybridization (aCGH). *Current Protocols in Cell Biology.*

Mondragón, R. J. (2017). Core-biased random walks in networks. *Journal of Complex Networks, 6(6),* 877-886.

Montenegro R. (2009). The simple random walk and max-degree walk on a directed graph. *Random Structures and Algorithms. 34(3),* 395-407.

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Bruce Spiegelman, B., Lander, E. S., Hirschhorn, J. H., Altshuler, D., & Groop, L. C. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics, 34(3),* 267–273.

Myslobodsky, M. (2008). Ingenuity Pathway Analysis of Clozapine-Induced Obesity. *Obesity Facts, 1(2),* 93-102.

Naithani, S., & Jaiswal, P. (2016). Pathway Analysis and Omics Data Visualization Using Pathway Genome Databases: FragariaCyc, a Case Study. *Methods in Molecular Biology Plant Genomics Databases,* 241-256.

Ning, Z., Feng, C., Song, C., Liu, W., Shang, D., Li, M., Wang, Q., Zhao, J., Liu, Y., Chen, J., Yu, X., Zhang, J., & Li, C., (2019). Topologically inferring active miRNA-mediated subpathways toward precise cancer classification by directed random walk. *Molecular Oncology, 13(10),* 2211–2226.

Odeh, A. (2017). Novel Genetic Algorithm for Early Prediction and Detection of Lung Cancer. *Journal of Cancer Treatment and Research. 5(2):*15.

Ong, H. F., Mustapha, N., & Sulaiman, M. N. (2011). Integrative Gene Selection for Classification of Microarray Data. *CIS Computer and Information Science, 4(2).*

Package igraph. (2019). Retrieved from https://cran.r project.org/web/packages/igraph /index.html.

Paszkiewicz, K., & Studholme, D. J. (2011). High-Throughput Sequencing Data Analysis Software: Current State and Future Developments. *Bioinformatics for High Throughput Sequencing*, 231-248.

Pawitan, Y., Bjöhle, J., Amler, L., Borg, A.-L., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., Liu, E. T., Miller, L., Nordgren, H., Ploner, P., Sandelin, K., Shaw, P. M., Smeds, J., Skoog, L., Wedrén, S., & Bergh, J. (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research, 7(6).*

Pemantle, R. (1988). Phase Transition in Reinforced Random Walk and RWRE on Trees. *The Annals of Probability, 16(3),* 1229-1241.

Petrochilos, D., Shojaie, A., Gennari, J., & Abernethy, N. (2013). Using random walks to identify cancer-associated modules in expression data. *BioData Mining, 6(1).*

Philipp, O., Osiewacz, H. and Koch, I. (2016). Path2PPI: An R package to predict protein–protein interaction networks for a set of proteins. *Bioinformatics. 32(9)*:1427-1429.

Polat, K., & Güneş, S. (2009). A new feature selection method on classification of medical datasets: Kernel F-score feature selection. *Expert Systems with Applications, 36(7),* 10367-10373.

Prete, E. D., Facchiano, A., & Liò, P. (2017). A Gene Set Enrichment Analysis of multiomic celiac disease data. *PeerJ Preprints.*

Rami-Porta, R., & Goldstraw, P. (2010). Strength and weakness of the new TNM classification for lung cancer. European Respiratory Journal, 36(2), 237-239.

Rastegar, R. (2012). Topics in self-interacting random walks. *Graduate Theses and Dissertations.*

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., and Bader, G. D. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols, 14(2),* 482-517.

Reimers, M. (2010). Making Informed Choices about Microarray Data Analysis. *PLoS Computational Biology, 6(5).*

Ren, G. and Liu, Z. (2012). NetCAD: a network analysis tool for coronary artery disease-associated PPI network. *Bioinformatics, 29(2),* 279-280.

Revathy N, Amalraj D. (2011). Accurate Cancer Classification Using Expressions of Very Few Genes. *International Journal of Computer Applications. 14(4):*19-22.

Reynolds, A. (2013). Selection pressures give composite correlated random walks Lévy walk characteristics. *Journal of Theoretical Biology, 332,* 117-122.

Richards, R. A. (2016). Biological classification: A philosophical introduction. New York: *Cambridge University Press.*

Roy, S., & Guzzi, P. H. (2015). Biological Network Inference from Microarray Data, Current Solutions, and Assessments. *Methods in Molecular Biology Microarray Data Analysis,* 155-167.

S. A. Gagliano, A. D. Paterson, M. E. Weale, and J. Knight. (2015). Assessing models for genetic prediction of complex traits: a comparison of visualization and quantitative methods. *BMC genomics, 16(1),* 405.

S. Cockwell (2011). Gene Set Enrichment Analysis. Retrieved June 10, 2018, from http://bioinformatics.knowledgeblog.org/2011/06/20/gene-set-enrichment-analysis/

S. Kim, M. Kon, C. Delisi. (2012). Pathway-based classification of cancer subtypes. *Direct. 7*: 21.

S. Nacu, R. Critchley-Thorne, P. Lee, and S. Holmes. (2007). Gene expression network analysis and applications to immunology. *Bioinformatics, 23(7),* 850-858.

Sarkar, D. (2019). Package lattice. Retrieved from https://cran.r-project.org/web/packages /lattice/index.html.

Sartor, M. A., Leikauf, G. D., & Medvedovic, M. (2008). LRpath: A logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics, 25(2),* 211-217.

Schauerhuber, M., Zeileis, A., Meyer, D., & Hornik, K. (2008). Benchmarking Open-Source Tree Learners in R/RWeka. Data Analysis, Machine Learning and Applications Studies in Classification, *Data Analysis, and Knowledge Organization,* 389–396.

Schliep K, Hechenbichler K. (2006). The kknn Package. Retrieved May 17, 2017, from ftp://ftp.auckland.ac.nz/pub/software/CRAN/doc/packages/kknn.pdf

Shchur, L., Heringa, J., & Blöte, H. (1997). Simulation of a directed random-walk model the effect of pseudo-random-number correlations. *Physica A: Statistical Mechanics and Its Applications, 241(3-4),* 579-592.

Shi Jing, L., Fathiah Muzaffar Shah, F., Saberi Mohamad, M., Moorthy, K., Deris, S., Zakaria, Z. and Napis, S. (2015). A Review on Bioinformatics Enrichment Analysis Tools Towards Functional Analysis of High Throughput Gene Set Data. *Current Proteomics. 12(1):*14-27.

Sootanan, P., Meechai, A., Prom-On, S., & Chan, J. H. (2011). Pathway-Based Microarray Analysis with Negatively Correlated Feature Sets for Disease Classification. *Neural Information Processing Lecture Notes in Computer Science*, 676-683.

Sootanan, P., Prom-On, S., Meechai, A., & Chan, J. H. (2012). Pathway-based microarray analysis for robust disease classification. *Neural Computing and Applications, 21(4).*

Sorlie, T. (2016). The Impact of Gene Expression Patterns in Breast Cancer. *Clinical Chemistry, 62(8),* 1150-1151.

Štefka D, Holeňa M. (2013). Performance of classification confidence measures in dynamic classifier systems. *Neural Network World. 23(4):*299-320.

Stöppler, M. C. (M.D.). (2019). 4 Types of Genetic Diseases - Symptoms, Causes & Human Genome. Retrieved May 10, 2019, from https://www.medicinenet.com/genetic_disease/article.htm

Subat, S., Mogushi, K., Yasen, M., Kohda, T., Ishikawa, Y., & Tanaka, H. (2018). Identification of genes and pathways, including the CXCL2 axis, altered by DNA methylation in hepatocellular carcinoma. *Journal of Cancer Research and Clinical Oncology, 145(3),* 675-684.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences, 102(43),* 15545–15550.

Tekade, R., & Rajeswari, K. (2018). Lung Cancer Detection and Classification Using Deep Learning. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).*

Tripathi, A., Venugopalan, S., & West, D. B. (2010). A short constructive proof of the Erdős–Gallai characterization of graphic lists. *Discrete Mathematics, 310(4),* 843-844.

Tsuchiya, M., Parker, J. S., Kono, H., Matsuda, M., Fujii, H., & Rusyn, I. (2010). Gene expression in nontumoral liver tissue and recurrence-free survival in hepatitis C virus-positive hepatocellular carcinoma. *Molecular Cancer, 9(1),* 74.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., & Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-

dimensional cancer genomics data using PARADIGM. *Bioinformatics, 26(12),* i237–i245.

Velsher, L. (2003). Genetic issues in the care of the adolescent patient. *Paediatrics & Child Health,8(1),* 36-39.

Vogl, G. W., Weiss, B. A., & Helu, M. (2016). A review of diagnostic and prognostic capabilities and best practices for manufacturing. *Journal of Intelligent Manufacturing, 30(1),* 79-95.

Wang, K., Li, M., & Bucan, M. (2007). Pathway-Based Approaches for Analysis of Genomewide Association Studies. *The American Journal of Human Genetics, 81(6).*

Wang, W., & Liu, W. (2018). Integration of gene interaction information into a reweighted random survival forest approach for accurate survival prediction and survival biomarker discovery. *Scientific Reports, 8(1).*

Wang, X., & Simon, R. (2011). Microarray-based cancer prediction using single genes. *BMC Bioinformatics, 12(1).*

Wang, X., Dalkic, E., Wu, M., & Chan, C. (2008). Gene module level analysis: Identification to networks and dynamics. *Current Opinion in Biotechnology, 19(5),* 482-491.

Wang, Y. (2017). Transcriptional Regulatory Network Analysis for Gastric Cancer Based on mRNA Microarray. *Pathology & Oncology Research, 23(4),* 785-791.

Wei, Z., & Li, H. (2006). Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics, 8(2),* 265-284.

Wood, A., Shpilrain, V., Najarian, K., & Kahrobaei, D. (2019). Private naive bayes classification of personal biomedical data: Application in cancer data analysis. *Computers in Biology and Medicine,105,* 144-150.

Wu, J. (2017). Feature Selection for Cancer Classification Using Microarray Gene Expression Data. *Biostatistics and Biometrics Open Access Journal. 1(2).*

Yang, Q., Wang, S., Dai, E., Zhou, S., Liu, D., Liu, H., Meng, Q., Jiang, B., & Jiang, W. (2017). Pathway enrichment analysis approach based on topological structure and updated annotation of pathway. *Briefings in Bioinformatics, 20(1),* 168–177.

Yang, S., & Naiman, D. Q. (2014). Multiclass cancer classification based on gene expression comparison. *Statistical Applications in Genetics and Molecular Biology, 0(0).*

Yossi A., Andrei Z. B., Anna R. K., Nathan L., and Steven P., (1996). Biased random walks. *Combinatorica, 16(1),* 1-18.

Yu, K., Ganesan, K., Tan, L. K., Laban, M., Wu, J., Zhao, X. D., Li, H., Carol, H. W. L., Zhu, Y., Chia, L. W., Hooi, S. C., Miller, L., & Tan, P. (2008). A Precisely Regulated Gene Expression Cassette Potently Modulates Metastasis and Survival in Multiple Solid Cancers. *PLoS Genetics, 4(7).*

Zhang, A., Lu, H., Wen, D., Sun, J., Du, J., Wang, X., Gu. W., & Jiang, J. (2018). The potential roles of long non-coding RNAs in lipopolysaccharide-induced human peripheral blood mononuclear cells as determined by microarray analysis. *FEBS Open Bio, 9(1),* 148-158.

Zhang, Q.-L., Zhang, G.-L., Xiong, Y., Li, H.-W., Guo, J., Wang, F., Deng, .X-Y., Chen, J.-Y., Wang, Y.-J., & Lin, L.-B. (2019). Genome-wide gene expression analysis reveals novel insights into the response to nitrite stress in gills of Branchiostoma belcheri. *Chemosphere, 218,* 609–615.

Zhao, F., Ge, Y.-Z., Zhou, L.-H., Xu, L.-W., Xu, Z., Ping, W.-W., Wang, M., Zhou, C.-C., Wu, R., & Jia, R.-P. (2017). Identification of hub miRNA biomarkers for bladder cancer by weighted gene coexpression network analysis. *OncoTargets and Therapy, 10,* 5551–5559.

Zhou, X., & Tuck, D. P. (2007). MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. *Bioinformatics, 23(9),* 1106-1114.

Zia, A., & Rashid, S. (2020). Systems Biology and Integrated Computational Methods for Cancer-Associated Mutation Analysis. *'Essentials of Cancer Genomic, Computational Approaches and Precision Medicine,* 335-362.